CREATING STATISTICAL RESOURCES FROM REAL DATASETS – THE STARS PROJECT

Penelope Bidgood
Kingston University, United Kingdom
bidgood@kingston.ac.uk

The aims of the STARS (Statistical Resources from Real Datasets) project are to make available real datasets and associated scenarios applicable to a range of disciplines and to develop learning and assessment materials to accompany these datasets for use with various packages. The project team, based in 4 universities in England, have developed worksheets in Psychology, Health and Business, using mainly Excel, MINITAB, SPSS *in both pdf format and* Word*. The worksheets are designed to be used in introductory statistics courses in service teaching and cater for a range of student abilities, backgrounds and needs. Further, resources for individualised datasets and assignments with solutions to be generated from the datasets have been produced. The materials developed and the concepts behind them have a far greater potential and use throughout the statistics teaching community.*

BACKGROUND

The STARS (Creation of Statistical Resources from Real Data) project is funded by the Higher Education Funding Council for England (HEFCE) through phase 4 of the Fund for the Development of Teaching and Learning (FDTL4). The aim of this major funding programme was to stimulate developments in learning and teaching and to secure wide up-take and implementation of good practice. This followed the review of learning and teaching quality in several subjects at universities in England and Northern Ireland during 1998-2000. Mathematics, Statistics and Operational Research (MSOR) was one of the subjects reviewed at a total of 71 institutions. University departments that had achieved high grades in the assessment exercise were invited to bid for funds. MSOR was in competition with 12 other discipline areas for funding; STARS was the only successful project in statistics, although there were two others in Mathematics.

The MSOR Overview Report (2000), which is based on the individual 71 subject review reports for each institution, provided a snapshot of the state of health of teaching and learning in the discipline. It noted that “Student engagement and performance has often been greatest when dealing with well-focused problems of a practical nature” and that “In some instances there is a dedicated mathematics base room with particular local resources such as study packs. Generally these facilities were greatly appreciated by the students.” Each of these points is addressed by the STARS project.

In addition, the project was informed by five regional workshops that were held in October-November, 2000, which had the aims of discussing with HE statistics and OR colleagues how the MSOR network could best stimulate the sharing of good practice and innovation in learning and teaching. The most useful workshops were those at which discussion centred on the needs and aspirations of HE colleagues and their departments. (Davies, 2001) At the workshops there were one or more representatives from 32 UK universities, who said they would like, amongst other things,

- Material to improve teaching and learning, especially to make it more dynamic;
- Web-based tutorial-type material;
- Examples of good practice in the use of remote access student learning;
- Good assessment practice;
- Examples of case studies or projects, including the corresponding data, and use with packages such as *MINITAB* and *Excel* in service courses;
- Sources of associated data;
- Connecting and involving the many teachers of statistics who are attached to non-mathematics and statistics departments;

- How to deal with (a) a wide range of mathematical backgrounds for beginning students and (b) large classes with a wide range of ability;
- Reach out and help in service teaching;
- Good service course examples, especially datasets with teaching tips;

Thus, the STARS project addresses issues highlighted in the MSOR overview report and the needs expressed by lecturers in the discipline.

INTRODUCTION

The STARS project began in October 2002 and is based in a consortium of 4 universities – Coventry, Kingston, Nottingham Trent and Oxford-Brookes. The 5 team members had a wealth of experience in developing teaching, learning and assessment materials, having taught, between them, a range of courses to students of varying backgrounds, abilities and needs. Further, the team had expertise in statistical computing and the application of statistics to a wide range of disciplines.

The project was involved primarily with the creation of learning resources whilst addressing *independent* learning. The goals were to make available real datasets, from accessible databases, in a form suitable for a learning and teaching resource in HE, construct learning materials to accompany these datasets, and develop such materials so that they can be used with various statistical packages for a range of student abilities, backgrounds and needs. The materials so produced would be accessible electronically to allow for hard-copy or web use by both staff and students whether on campus or at a distance.

In particular, the stated aims of the STARS project were

- To improve students' experience of learning statistics;
- To create a national learning resource of statistics material for use by lecturers across a range of disciplines;
- To ensure these resources are available to all relevant students, irrespective of circumstances;
- To address the problem of retention by inspiring students with relevant scenarios;
- To impact on the professional development of lecturers;
- To address the problem of copying/plagiarism of student assignments/case studies.

All of the 4 universities subscribed to these aims as they were in agreement with their own Learning and Teaching Strategies. In particular, a key objective of the (then) lead site was "To support the continuing development of high quality, responsive teaching, learning and assessment that recognises the diversity of learners, the appropriate use of learning technologies and the dispersed nature of the university." Moreover, since all materials reflect real, in-context scenarios, they address a number of issues, not least of which are staff development, whether lecturers be statisticians or not, and student motivation and retention.

DATASETS

One of the main challenges of the project was the identification of suitable, real datasets and the determination of the accessibility of the corresponding databases. There are many web sites that, taken together, allow thousands of data sets to be downloaded. Many of these datasets may have meaningful variable names but, rarely, a description of the context in which the data were obtained. Also, there are very few that provide help with teaching and learning activities using the data set. Moreover, many of the relevant sites are international, or even just North American, in emphasis, whereas it is felt that a UK context would make the worksheets more accessible and interesting to students.

When the data are real and interesting, students get extra motivation in working with them. Students want to use material which relates to them and their discipline – anything else is likely to switch them off. Consequently the materials produced should help motivate students. However, there is rarely sufficient time for the lecturer to create context examples in every module. The STARS project aims to address this dilemma by creating in-context resources and to make them widely available.

The project team decided, in response to what lecturers said they needed, that the initial focus would be on the production of worksheets for three service course disciplines. This would also aid undergraduate students who do not always see immediately the relevance of the statistics course to their own field. Health Sciences was the first service area that was tackled. Statistics is heavily used in this discipline, which often involves large classes. Moreover, previously, one of the team had developed study materials based on some real, available, accessible data in this field.

Many subject areas, including engineering, psychology, life sciences, business, medicine and pharmacy need statistics. Lecturers were asked, through an online questionnaire, in which subject area(s) would they most like materials and whether they had any data sets with the accompanying scenarios. Responses to this and from an Association of Statistics Lecturers in Universities (ASLU) conference, early in 2003, encouraged the team to select Psychology and Business as the other two areas for production of worksheets. Statistics is an integral part of these disciplines and is generally taught in large classes.

Further, a suggestion was made to look at the STEPS (*St*atistical *E*ducation through *P*roblem *S*olving) data sets. This had been a project to support the teaching of statistics in Biology, Business, Geography and Psychology through problem based modules, based in *MINITAB*. Unfortunately, generally, the materials were difficult to access. The STARS project was able, with permission from the STEPS team, to use 3 of their Psychology datasets. It was decided to have 5 datasets for each of the 3 discipline areas. The remaining datasets generally came from other, academic colleagues in the consortium universities and elsewhere, although a few came from industrial and business contacts.

THE WORKSHEETS

As well as access to real data, lecturers said that they wanted corresponding access to routine exercises derived from the data that would help them to get students better to learn a range of statistical topics. Consequently, one of the main aims of this project was to produce discipline-specific worksheets which could be accessed electronically, using different (statistical) packages and which could address special needs. The worksheets are intended to support, but not be a substitute for, lecture materials.

For the first Health Sciences datasets, worksheets were produced in *MINITAB*, *Excel*, *SPSS* and *SAS*. The latter was soon dropped, however, as the team decided to concentrate on introductory statistics courses and very few, if any, of these use *SAS*. Hence, all worksheets are available in the three main packages used in service teaching in universities in the UK and a few are also available in *SAS*. Although computing packages are continually updated, a decision was made not to change the worksheets once they had been written. As new versions are available, then the worksheets can be altered by lecturers to take account of this. The one exception to this “rule” was the decision to update the *MINITAB* worksheets when version 14 became available. The project team felt that this was a sufficient improvement on previous versions, particularly for graphical work, to make the update worthwhile.

Typically, each dataset has 3 or 4 associated worksheets; topics vary according to the data being investigated. As the worksheets are designed to be used in introductory statistics courses in service teaching, typical statistics topics covered are Charts, Descriptive Statistics, Tests of Means, Non-parametric Tests, Tests of Association, Correlation and Regression, Further Regression and Time Series. Examples of worksheet subjects for the 5 Health Sciences scenarios are shown in Table 1.

The worksheets are designed to cater for a range of student abilities, backgrounds and needs. Following advice from support services at the consortium universities, the worksheets were written in Arial as this is deemed to be easier to read for dyslexic students. Most worksheets contain a section on Further Investigation or Further Work to stretch the more-able student. All worksheets are available with suggested answers and they are designed to be able to be used by students individually at home, or in a class or tutorial setting.

Table 1: Statistical Areas covered in the Health Sciences

Dataset	Typical question	Statistical topic
Obesity	*Did patients receiving the new drug lose significantly more weight than those who received the placebo?*	Charts
		Descriptive Statistics
		Tests of Means
		Tests of Association
Triglyceride	*Can we predict change in triglyceride level from change in weight?*	Charts
		Descriptive Statistics
		Correlation and Regression
		Further Regression
Breakfast	*How does adding sugar to low GI breakfast affect children's hunger later in the day?*	Charts
		Descriptive Statistics
		Tests of Means
		Non-parametric Tests
Overdose	*Do gender, age or marital status affect the risk of taking an overdose?*	Charts
		Descriptive Statistics
		Tests of Association
IVF	*Are larger clinics more successful than smaller ones at IVF treatment?*	Descriptive Statistics
		Tests of Means
		Correlation and Regression

ASSESSMENT

Another main aim of the STARS project was to develop assessment tools that could produce individualised assignments, (with suggested solutions for lecturers) which could address the problem of plagiarism. Assessment had emerged as one of the most problematic areas in MSOR provision in the subject Overview Report (2000). Some of the main issues were deemed to be the link between what students are expected to be capable of (learning objectives) and the assessment tasks set; the match of the assessment task to the student profile; the management and administration of the design of the assessment; setting and moderation of assessment task and scheduling of assessment tasks. Further, good practice involved a wide range of assessment instruments to be used to address learning objectives (Bidgood and Cox, 2000).

The two assessment tools developed in this project help to address these issues; each allows individualised datasets to be produced, together with assignments (and solutions) that can be generated from the datasets. The primary motivation for creating individualised tasks for students is to combat plagiarism within the student group, where a student may copy all or part of another student's work (rather than the more usual form of plagiarism, i.e., copying material without reference from published sources). It is important to guarantee that each student, in carrying out statistical analyses, works with a different set of data.

Individualised Statistics Coursework Using Spreadsheets (ISCUS) is based on *Excel* and allows lecturers to use their own data as well as that supplied in the STARS project (Hunt, to appear). As such it facilitates the setting and marking of student assignments based on a substantial set of data. However, random data allocation can lead to problems, for example, in a regression problem, some students might have an outlier in their data and others not. An advantage of ISCUS is that, generally, lecturers can identify such problems in advance and ensure that the samples of data received are comparable in their statistical features. ISCUS has been developed over a period of time and, as such, has been introduced to the community through various workshops and conferences, as part of the dissemination strategy of the project. It is freely downloadable from the website, http://stars.ac.uk, together with instructions for its use.

The other assessment tool is not tied to any package but is geared to be used with the STARS datasets and operates over the Internet. Here, a browser-enabled worksheet creator has been developed that uses a version of RWeb to add value to the data set by allowing the creation of unique worksheets with solutions, based on random samples of values of some of the variables in the database. Lecturers can generate as many unique worksheets and solutions as they wish and these can be disseminated to students electronically or using paper-based methods. Students are able to access the facility from a web browser and practice their data analytic skills as often as they like, either at times when they have access to the Internet, or when they can work through paper-based versions in the traditional way (Davies and Payne, 2001). It is clear that such facilities as these assessment tools are useful to both teachers and students of statistics.

EVALUATION

Evaluation of the project has taken various forms. The worksheets, for example, have been monitored by the project team themselves, by colleagues who teach statistics in other universities and by students in statistics classes. Opportunities to disseminate the work, including the assessment tools, have been afforded by members of the team attending various workshops and conferences throughout the 3 years that the project has run. The project, as a whole, has an external moderator, as required by HEFCE.

The project team developed a protocol for evaluating the worksheets between themselves. Each data set was "adopted" by one of the team, who produced 3 or 4 worksheets, based on the data, in one of the 3 packages, *Excel*, *MINITAB* or *SPSS*. These worksheets were then sent to a designated team member for checking and evaluation and returned with comments. A revised set was produced (usually after some discussion between the team members), which were then ready for other team members to rewrite, where possible, in the other packages. Of course, limitations of the packages meant that not all worksheets could be reproduced exactly. Final versions were sent to the project director to check house style and load onto the web.

One of the first evaluation methods was a set of questionnaires asking for feedback from other statistics lecturers on some of the early worksheets. Questions involved the worksheet content, format and flexibility; the responses were overwhelmingly positive. The questionnaire also asked for suggested improvement and modifications to the worksheets and where they might be used. This gave the most valuable feedback – for example about updating to *MINITAB 14*, a suggestion that was readily adopted by the team. Again, most comments were positive, praising the use of real data ("The example I looked at would really motivate our students") and the format ("I particularly like the way that the text is designed to help the student to stop and think about what they are doing and what the analyses are showing them"). There were differing views on the ability to tailor the worksheets for local use, from "being able to tailor the worksheet is one of the strengths of this project" to "it's excellent as it is."

Of course, the real test is how students react to the worksheets. Initial piloting was carried out in two Business Studies classes, both using *Excel*-based worksheets based on data about the fast food market. Two-hundred females and 200 males were asked about their fast food purchasing – for example, how often they bought fast food and what brand they bought last- as well as questions about themselves (e.g., age, gender). The evaluator (external to the project team) observed the classes and talked to the students afterwards. It was observed that there was a great difference in the times students took to complete the worksheets, ranging from 10 minutes to 2 hours and that students discussed their work very little, except to check their answers. In the follow-up discussions, students liked the scenarios; found the instructions easy to follow and enjoyed "doing the statistics" and filling in the answers.

CONCLUSIONS

There is a popular view abroad that there is a vast amount of UK-based university-level statistics teaching material freely available on the internet for general use. A recent review by Barnett (2004) concluded that this is probably not so. The STARS project goes some way to redress the balance. All of the materials- scenarios, datasets worksheets and assessment tools- are available freely on a dedicated website http://stars.ac.uk which is hosted by Glasgow University.

The project has increased awareness of the value of statistics by providing a catalogue of real datasets, with discipline specific worksheets, based on real scenarios and using a range of software, available over the web. The worksheets can be accessed via Adobe Acrobat and Microsoft Office, the latter enabling modification by staff for local use, and both addressing special needs requirements. Further, procedures for generating random subsets of data so that students can have individualised datasets and assignments with suggested solutions, have been developed.

This project has not been able to cover all the subjects which could benefit from the creation of statistical materials of the form considered here. It is hoped that the teaching, learning and assessment opportunities afforded by the worksheets and assessment tools developed within the STARS project will be of benefit to the statistics teaching community, especially in service teaching classes. Although STARS was initiated by the needs of staff in the UK and funded as a result of a quality assurance exercise in England and Northern Ireland, the materials developed and the concepts behind them have a far greater potential and use throughout the statistics teaching community.

REFERENCES

Barnett, V. (2004). Review of online statistics teaching material. *MSOR Connections*, 4(2), 43-45.

Bidgood, P. and Cox, W. (2002). Student assessment in MSOR. *MSOR Connections*, 2(4), 9-13.

Davies N. (2001). Feedback from the Stats/OR community. *MSOR Connections*, 1, 3-4.

Davies, N. and Payne, B. (2001). Web-created real data worksheets. *MSOR Connections*, 1(4), 15-17.

Hunt, N. (to appear). Individualised statistics coursework using spreadsheets *Teaching Statistics*.

Mathematics, Statistics and Operational Research Overview Report. (2000). Gloucester, UK: Quality Assurance Agency for Higher Education.